Metrics for Evaluation of Ontology-based Information Extraction

Diana Maynard
Dept of Computer Science
University of Sheffield
211 Portobello St
Sheffield, UK
diana@dcs.shef.ac.uk

Wim Peters
Dept of Computer Science
University of Sheffield
211 Portobello St
Sheffield, UK
wim@dcs.shef.ac.uk

Yaoyong Li
Dept of Computer Science
University of Sheffield
211 Portobello St
Sheffield, UK
y.li@dcs.shef.ac.uk

ABSTRACT

The evaluation of the quality of ontological classification is an important part of semantic web technology. Because this area is under constant development, it requires improvement and standardisation. This paper discusses existing evaluation metrics, and proposes a new method for evaluating the ontology population task, which is general enough to be used in a variety of situations, yet more precise than many current metrics. The paper further describes our first efforts in operationalising the evaluation procedure, including the creation of a semantically annotated corpus that will function as a test bed for the proposed evaluation mechanism, and comparison of different evaluation metrics. We conclude that for ontology-based evaluation, a more complex mechanism than is traditionally used is preferable. This mechanism aims to drive a benchmarking assessment tool for the current state-of-the-art of ontology population, and to set a standard for best practice for future evaluation of human language technology for the semantic web.

1. INTRODUCTION

Natural language techniques involving the classification of text by means of ontologies are a relatively new area of research, and while they are mainly derived from applications (and their respective evaluation methods) with a long history, methods for evaluation of such technologies are currently at the forefront of research. As yet, no widely accepted standards have emerged.

The automatic assignment of ontological classes to text elements is an important aspect of semantic web applications and services. It can be postulated that the more successful this assignment is, the more feasible the ontology is for semantic web applications in terms of reproducible semantic indexing. The outcome of this process may, in a number of cases, give an indication of the actual structural and conceptual quality of the ontology involved. For instance, consistent automatic misclassification may flag a wrongly positioned concept. Multiple classifications (i.e. where the system consistently assigns more than one class to the same text element) may point to a lack of distinguishability between concepts and may ultimately suggest either a merging of such concepts or further explicit specification to enhance their distinguishability. Overall, automatic classification may yield valuable clues for the evaluation of ontologies.

Because the development of semantic web tools and resources is currently hampered by its very diversity, it is important to have some standardised metrics and evaluation methods, in order to promote development of tools in a productive rather than random way, and to enable end users to make informed choices about the best tools for their needs.

In this paper we discuss the specifications for a proper evaluation of ontology population, describe current techniques, and examine to what extent existing methodologies can be reused in the context of ontology population. We then propose an evaluation method, which takes inspiration from several existing methodologies. We describe the creation of a text corpus enriched with ontological information from a freely available general purpose ontology, which can be used as a gold standard for evaluating state-of-the-art annotation systems. Finally we describe some experiments with different evaluation metrics using this corpus and ontology.

2. DEFINING PERFORMANCE METRICS

In order to compare the quality of ontology-based annotation tools and other HLT applications associated with the semantic web, we need some evaluation metric which can provide a simple mechanism for comparing different systems and different versions of the same system in a consistent and repeatable way. Ideally, the same metrics should be used by everyone, so that comparisons can be made between different evaluations. This requires a simple but powerful evaluation metric that can be easily implemented by other people, and/or a freely available tool that offers such functionality.

In this section, we shall examine the case of ontology-based information extraction (OBIE), which is used as the basis for automatic semantic annotation / metadata extraction. This is an important component of the semantic web, since ontologies must be populated with information from documents, and documents need to be semantically annotated. Currently there is no standard for OBIE, because it is a relatively new area of research, although as will be discussed in Section 3, there are several well-established metrics for evaluation of traditional information extraction systems. The needs of OBIE metrics are rather different, however, because traditional methods are binary rather than scalar. This means that they assess an answer as correct or incorrect (occasionally allowing for partial correctness which is generally allocated a "half-score"). Ontology-based systems

WWW2006, May 22–26, 2006, Edinburgh, UK.
.

should, however, be evaluated in a scalar way, in order to allow for different degrees of correctness. For example, a scalar method allows the score to be based on the position of the response in the ontology and its closeness to the correct position in the ontology.

The evaluation task for ontology-based information extraction aims to discover in the text all mentions of instances related to the ontology. The gold standard is a set of texts where instances are annotated with their related ontological concepts. We aim to measure how good the IE system is at discovering all the mentions of these instances, and whether the correct class has been assigned to each mention.

When defining the metric, we suggest the following criteria as proposed by [15]. The metrics should: reach their highest value for perfect quality; reach their lowest value for worst possible quality; be monotonic; be clear and intuitive; correlate well with human judgment; be reliable and exhibit as little variance as possible; be cheap to set up and apply; be automatic.

3. EXISTING EVALUATION METRICS

In this section, we describe the main existing metrics for evaluation of IE and related tasks, and discuss their suitability for use or adaptation to ontology-based tasks.

3.1 Precision and Recall

Many human language technology (HLT) tasks such as Information Extraction are traditionally evaluated using Precision, Recall and F-measure. These metrics have a very long-standing tradition in the field of IR [26, 20, 10]). For example, they have been used in large-scale IE evaluations such as MUC (Message Understanding Conferences)[3] and CONLL [25, 7]. Because much of the research in IE in the last decade has been connected with these competitions, and because of the availability of the gold standard corpora used for them, it has become natural for people to compare their systems on this data and with the same metrics, which means that they have been the most widely used in this field, albeit with slight variations from time to time.

Precision measures the number of correctly identified items as a percentage of the number of items identified. In other words, it measures how many of the items that the system identified were actually correct, regardless of whether it also failed to retrieve correct items. The higher the Precision, the better the system is at ensuring that what has been identified is correct. It is formally defined as

$$Precision = \frac{Correct + 1/2 Partial}{Correct + Spurious + Partial} \tag{1}$$

Note that we consider annotations to be partially correct if the entity type is correct and the spans are overlapping but not identical. Partially correct responses are normally allocated a half weight.

Recall measures the number of correctly identified items as a percentage of the total number of correct items. In other words, it measures how many of the items that should have been identified actually were identified, regardless of how many spurious identifications were made. The higher the Recall rate, the better the system is at not missing correct items. Recall is formally defined as:

$$Recall = \frac{Correct + 1/2 Partial}{Correct + Missing + Partial} \tag{2}$$

The **F-measure** [26] is often used in conjunction with Precision and Recall, as a weighted average of the two. If the weight is set to 0.5 (which is usually the case), Precision and Recall are deemed equally important. F-measure is formally defined as:

$$F - measure = \frac{(\beta^2 + 1) P * R}{(\beta^2 R) + P} \tag{3}$$

where β reflects the weighting of P vs. R. If P and R are to be given equal weights, then we can use the equation:

$$F_1 = \frac{P * R}{0.5 * (P + R)} \tag{4}$$

3.2 Cost-based Evaluation Metric

False positives are also a useful metric when dealing with a wide variety of text types, because they are not dependent on relative document richness. By this we mean the relative number of entities or annotations of each type to be found in a set of documents. When comparing different systems on the same document set, relative document richness is unimportant, because it is equal for all systems. When comparing a single system's performance on different documents, however, it is much more crucial, because if a particular document type has a significantly different number of any type of entity, the results for that entity type can become skewed. This could be used to manipulate a system's reported results by testing on a particular corpus (the results would be accurate, but misleading). Compare the impact on Precision of one error where the total number of correct entities = 1, and one error where the total = 100. Assuming the document length is the same, then the false positive score for each text, on the other hand, should be identical. The False Positive metric is formally defined as:

$$FalsePositive = \frac{Spurious}{c} \tag{5}$$

where c is some constant independent from document richness, e.g. the number of tokens or sentences in the document.

Another solution is to use error rate, which is the inverse of Precision, and measures the number of incorrectly identified items as a percentage of the items identified. It has the advantage over False Positives that it does not require some arbitrary constant, which can make the results hard to interpret, and means that comparison on documents of different length is also skewed.

Using error rate instead of Precision and Recall means, however, that the F-measure can no longer be used. An alternative method of getting a single bottom-line number to measure performance is the cost-based evaluation (CBE) metric. This has been used in some of the DARPA competitions, such as TDT2 [9], and ACE [1]. The model stems from the field of economics, where the standard model "Time Saved x Salary" measures the use of the direct salary cost to an organisation as a measure of the value.

One of the main advantages of this method is that it enables the evaluation to be adapted depending on the user's requirements, and so is particularly suitable for use in industry where one wants to choose between different systems for a particular task. A CBE model characterises the performance in terms of the cost of the errors. For any application, the relevant cost model is applied, and expected prior target statistics are defined.

For a cost-based error model, a cost would typically be associated with a miss and a false alarm (spurious answer), and with each category of result (e.g. recognising Person might be more important then recognising Date correctly). Expected costs of error would typically be based on probability (using a test corpus). This makes the assumption that a suitable test corpus is available, which has the same rate of entity occurrence (or is similar in content) to the evaluation corpus. If necessary, the final score can be normalised to produce a figure between 0 and 1, where 1 is a perfect score.

3.3 Learning Accuracy

Another solution consists of augmenting the traditional Precision and Recall metrics by adding some kind of semantic distance weights, such that the gravity of the error can be taken into account.

Cimiano et al [4] use a method called Learning Accuracy to evaluate how well an ontology has been populated. This was originally used by Hahn et al [13] to measure how well a concept had been added in the right level of the ontology, but it can be equally applied to measure how well the instance has been added in the right place. Learning Accuracy (LA) essentially measures "the degree to which the system correctly predicts the concept class which subsumes the target concept to be learned".

LA uses the following measurements:

- SP (Shortest Path) = the shortest length from root to the key concept
- FP = shortest length from root to the predicted concept. If the predicted concept is correct, then FP = 0, i.e. FP is only considered in the case that the answer given by the system is wrong.
- CP (Common Path) = shortest length from root to the MSCA (Most Specific Common Abstraction, i.e.the lowest concept common to SP and FP paths)
- DP = shortest length from MSCA to predicted concept

If the predicted concept is correct, i.e. if FP =0,

$$LA = \frac{CP}{SP} = 1 \tag{6}$$

If the predicted concept is incorrect,

$$LA = \frac{CP}{FP + DP} \tag{7}$$

Essentially, this measure provides a score somewhere between 0 and 1 for any concepts identified in an incorrect position in the ontology. If a concept is missing or spurious, the score is 0, and if it is correct, the score is 1 (as with Precision and Recall). So this method provides an indication of how serious the error is, and weights it accordingly.

3.4 Augmented Precision and Recall

Augmented Precision and Recall, the evaluation metric we propose in this section, aims to preserve the useful properties of the Precision and Recall scoring metrics, but combine them with a cost-based component.

The Precision/Recall model is the most well known and widely used evaluation metric in the IE community, though this does not automatically make it the most suitable for all tasks. It has a big advantage over metrics such as CBE in that it is easily understood. It can be difficult to interpret the CBE results in a meaningful way, partly because the error costs can be changed for each application, making it difficult to compare systems unless the same costs are used, and partly because the single-figure score generated is not as meaningful or easily understood as the percentage score given by the F-measure.

Another problem with the CBE metric is that it contains many different costs, which are assigned by the end-user, and it is not easy to decide on appropriate weights or to find a way to calculate these automatically. One method is to initially make all weights the same, and include a distance-based metric for ensuring that partially correct items, which are assigned a tag at the wrong level of the ontology, are penalised appropriately according to the distance. The consequence of this is that the weights become superfluous for the task of automatically evaluating ontology population without a preliminary stage of complicated and subjective weight setting.

We do need to stress the superiority of the CBE model over Precision and Recall in another respect: that it allows multi-dimensional evaluation, where a single score is not generated, but instead the evaluation is carried out simultaneously along several axes. This model is designed specifically for different applications or different users, who might have diverging requirements of a system. For example, one user might be more concerned with Precision than Recall, or one user might be more concerned about getting particular types of entities right, and not so concerned about other types, or one user might be more concerned with the fact that getting something partially right is important. Therefore a cost-based model is useful because it enables the parameters to be modified according to the particular evaluation or task.

Multi-dimensional evaluation has been applied in several systems. For example, Olsson et al. [22] evaluate the performance of protein name taggers in this way to overcome the limitations of Precision and Recall being too inflexible, proposing additional measures such as Sloppy, Left Boundary and Right Boundary to cater for responses which overlap the Key annotations. The GATE evaluation tools [6] provide something similar, where partially correct answers can be given a half weight (Average), counted as correct (Lenient) or counted as incorrect (Strict).

However, if a fully-fledged CBE model were to be adopted as a standard for ontology population evaluation, we would have to devise some simple and heuristic method of weight assignment, or in any case the creation of a generic set of weights that could be used as a default. Also, we would need some scoring tool, with the ability to be adapted easily by the user to reflect changes to the weights. Although the CBE model guarantees the most flexible application of various evaluation metrics, we have opted for a simpler version where we only take two dimensions into account: the Precision/Recall metric, and semantic distance between key (gold standard) and response (system output) concepts in terms of a given ontology (similar to TRUCKS[21] and LA). This method measures how well a particular text item has been classified.

One of the problems with LA is that it does not take into account the depth of the key concept in the hierarchy, considering essentially only the height of the MSCA and the distance from the response to the MSCA. This means that however far away the key is from the MSCA, the metric will give the same outcome. It could be argued that this is not important for similarity. However, this means that similarity ceases to be bi-directional, in that similarity between two concepts differs according to which is the key and which is the response. We shall see examples of this later in Section 4.2, which clearly look counter-intuitive to human judgement.

We therefore propose a more balanced distance metric, which we call BDM. This uses the following measurements:

- MSCA: most specific concept common to the key and response paths
- CP: shortest path from root concept to MSCA
- DPR: shortest path from MSCA to response concept
- DPK: shortest path from MSCA to key concept

Note that we maintain the labelling system used by LA, although some of the labels are not very intuitive, extending DP to DPR (Response) and DPK (Key) to make it clearer. Each of the paths has been normalized with two additional measurements, of which the first is the average length of the chains in which key and response concepts occur. The longer a particular ontological chain is, the more difficult it is to consistently pick out a particular class for annotation [2]. The second normalization is the introduction of the branching factor (i.e. number of descendants) of the relevant nodes in the ontology. This is also an indication of the level of difficulty associated with the selection of a particular ontlogical class relative to the size of the set of candidates. These normalizations will make the penalty that is computed in terms of node traversal within our metric relative to the semantic density of the chains.

The following concrete implementations are the result:

- n0: the average chain length of the whole ontology, computed from the root concept.
- n2: the average length of all the chains containing the key concept, computed from the root concept.
- n3: the average length of all the chains containing the response concept, computed from the root concept.
- BR: the branching factor of each relevant concept, divided by the average branching factor of all the nodes from the ontology, excluding leaf nodes.

The complete BDM formula is as follows:

$$BDM = \frac{BR(CP/n0)}{BR(CP/n0) + (DPK/n2) + (DPR/n3)} \qquad (8)$$

This measure takes the relative specificity of the taxonomic positions of the key and response into account in the score, but it does not distinguish between the specificity of the key concept on the one hand, and the specificity of the response concept on the other. For instance, the key can be a specific concept (e.g. 'car'), whereas the response can be a general concept (e.g. 'relation'), or vice versa. It is conceivable that a system's performance is evaluated differently in either situation, although a score along the dimension of the difference in relative key and response generality does not seem intuitively straightforward. If we wanted to integrate this aspect into the metric, we could add the specificity comparison of key and response concepts as a separate weight to the formula.

Essentially, our measure provides a score somewhere between 0 and 1 for the comparison of key and response concepts with respect to a given ontology. If a concept is missing or spurious, BDM is not calculated since there is no MSCA. If the key and response concepts are identical, the score is 1 (as with Precision and Recall). Overall, in case of an ontological mismatch, this method provides an indication of how serious the error is, and weights it accordingly.

We can now combine the BDM scores for each instance in the corpus, to produce Augmented Precision, Recall and F-measure scores for the annotated corpus. We differentiate this from traditional Precision and Recall due to the fact that it considers weighted semantic distance in addition to a binary notion of correctness.

$$\begin{aligned} BDM &= \sum_{i=\{1\dots n\}} BDM_i \\ BDM_i &= \frac{BR(CP_i/n0)}{BR(CP_i/n0) + (DPK_i/n2_i) + (DPR_i/n3_i)} \end{aligned} \qquad (9)$$

Augmented Precision (AP) and Recall (AR) for the corpus are then calculated as follows:

$$\begin{aligned} AP &= \frac{BDM}{BDM + Spurious} \\ AR &= \frac{BDM}{BDM + Missing} \end{aligned} \qquad (10)$$

while F-measure is calculated from Augmented Precision and Recall as:

$$F_1 = \frac{AP * AR}{0.5 * (AP + AR)} \qquad (11)$$

Note that we could replace BDM by another metric such as LA in Equation 10 if we wish. This is used in our experiments in Section 4.2 in order to compare LA and BDM on equal terms.

4. EVALUATION PROCEDURE

In order to enable the application and evaluation of the evaluation algorithms proposed in the previous section, we need an ontology and a text corpus that is semantically annotated with concepts from the ontology. In general, this method works if both resources are available for a particular conceptualisation, expressed in the ontology, and corresponding text annotations. A restriction on the nature of the ontology is that it must include hierarchical chains. For the evaluation matrix to be effective it cannot be just a set of named entities with no taxonomic embedding. If named entities are used as the only evaluation criterion, a binary metric with standard Precision and Recall suffices, i.e. the evaluation is in that case based on (partial) matching, missing annotations and false positives. To evaluate conceptual matching with respect to an ontology, we require a more complex evaluation mechanism such as LA or Augmented Precision and Recall, as described above. These metrics also fulfill important evaluation criteria such as ease of implementation, simplicity, coverage, scalability, repeatability, and ease of comprehension of the results [11].

4.1 The OntoNews Corpus

For our evaluation, we have created a corpus with semantic annotation[23], which is to be used as a gold standard. This is known as the OntoNews corpus. This semantically annotated corpus consists of 292 news articles from three news agencies: The Guardian, The Independent and The Financial Times. The news articles cover the period of August to October, 2001. The articles belong to three general topics or domains of news gathering: International politics, UK politics and Business.

The ontology used in the generation of the ontological annotation process is the PROTON ontology[1], which has been created and used in the scope of the KIM platform[2] for semantic annotation, indexing, and retrieval [16]. The PROTON ontology forms part of an annotation tool for automatic ontology population and open-domain dynamic semantic annotation of unstructured and semi-structured content for Semantic Web knowledge management applications. The ontology consists of around 250 classes and 100 relations. PROTON has a number of important properties, e.g. it is domain-independent, and therefore suitable for the news domain, and it is modular (comprising both a top ontology and a more specific ontology).

The overall objective of annotation was to manually create a gold standard with a high level of annotated knowledge contained within the PROTON ontology. The result is an annotation set that should be able to cover a variety of levels and types of semantic annotation, and is decomposable into subsets that constitute three different types of concepts:

1. **Named entities**: a small set of semantic classes which can be readily chosen from the PROTON ontology. Typically, this set includes people, organizations, locations, facilities, geo-political entities and time expressions. They are mostly referred to by proper names. The most well known named entity competitions in which annotations have been produced together with evaluation procedures of automatic annotations are MUC[12] and ACE[3].

2. **Top ontology**: a subset of PROTON with 20 high level concepts. This set contains the three unique beginners (top level concepts) from PROTON, together with a number of their high level hyponyms:

   - Abstract with hyponyms such as BusinessAbstraction and SocialAbstraction;
   - Object with direct hyponyms such as InformationResource, Agent and Organization;
   - Happening with direct hyponyms such as Situation, Event and TimeInterval.

3. **Concepts denotated by common nouns**. This type of annotation takes semantic coverage beyond that of proper names, and concentrates on the annotation of common nouns as they appear in the text.

   The advantage of extending the scope of semantic annotation to common nouns is that the result is a much more detailed and varied semantic characterisation of the domain involved and the entities that play a significant role in it. This extra information is necessary for more fine-grained semantic processing tasks.

Overall, the use of PROTON as a more or less fully-fledged ontology in our annotation significantly extends the semantic coverage of our annotation compared with previous and ongoing initiatives. The semantics of the annotation is more complex than initiatives such as MUC and ACE because of the relative fine-grainedness of the PROTON ontology, with hierarchical chains containing up to eight nodes.

Figure 1 shows a sample document from the OntoNews corpus annotated in GATE with instances from the PROTON ontology. In the main window is shown the text annotated by means of colour coding. On the right is a section of the ontology. Details of the instances and their relation to the ontology are also included, but are not shown in this screenshot.

4.2 Experiments

The aim of the experiments carried out on the OntoNews corpus was, on the one hand, to evaluate a new learning algorithm for OBIE, and, on the other hand, to compare the different evaluation metrics.

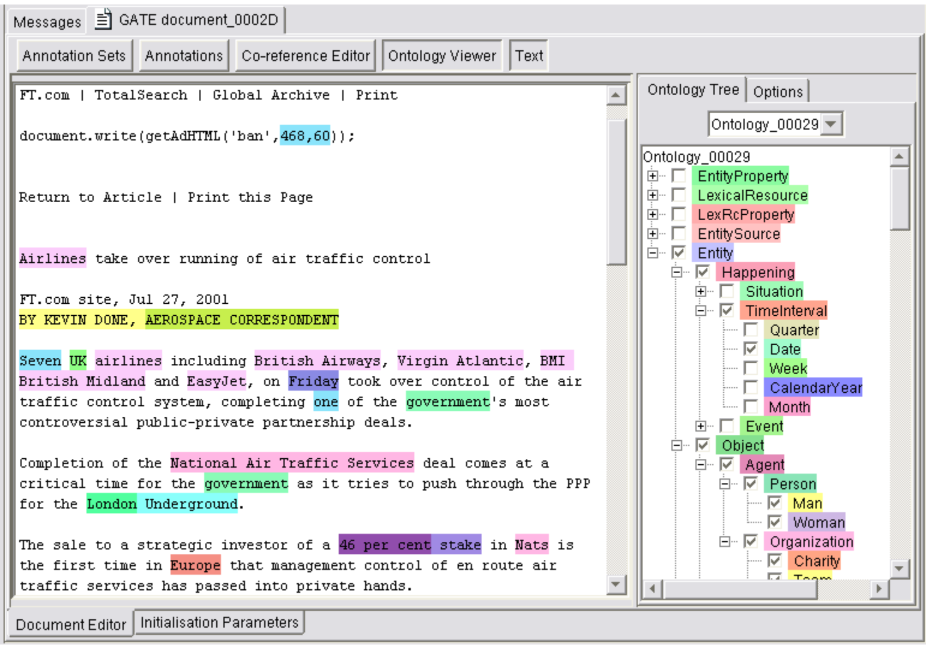
The OBIE algorithm learns a Perceptron classifier for each concept in the ontology; meanwhile it tries to keep the difference between two classifiers proportional to the cost of their corresponding concepts in the ontology. In other words, the learning algorithm tries to classify an instance as correctly as it can. If it cannot classify the instance correctly, it then tries to classify it with another concept with the least cost associated with it relative to the correct concept. The algorithm is based on the Hieron, a large margin algorithm for hierarchical classification proposed in [8]. See [19] for details about the learning algorithm and experiments.

We experimentally compared the Hieron algorithm with the learning algorithm SVM (see e.g. [5]) for OBIE. The SVM is a state of the art algorithm for classification. [17] applied the SVM with uneven margins, a variant of the SVM, to the traditional information extraction problem and achieved state of the art results on several benchmarking corpora. In the application of the SVM to OBIE, we learned one SVM classifier for each concept in the ontology separately and did not take into account the structure of the ontology. In other words, the SVM based IE learning algorithm was a flat classification in which the structure of concepts in the ontology was ignored. In contrast, the Hieron algorithm for IE was based on hierarchical classification that exploits the structure of concepts.

As the OntoNews corpus consists of three parts (International politics, UK politics and Business), for each learning algorithm two parts were used as training data and another part as test data. Note that although the tripartition of the corpus indicates three distinct and topically homogeneous parts of the corpus, these parts are used as training and testing data for the comparison of different algorithms, and not their performance. For this purpose, semantic homogeneity does not play a role.

For each experiment we computed three F_1 values to measure the overall performance of the learning algorithm. One was the conventional micro-averaged F_1 in which a binary

[1] http://proton.semanticweb.org
[2] http://www.ontotext.com/kim
[3] http://www.nist.gov/speech/tests/ace/index.htm



```
FT.com | TotalSearch | Global Archive | Print

document.write(getAdHTML('ban',468,60));


Return to Article | Print this Page


Airlines take over running of air traffic control

FT.com site, Jul 27, 2001
BY KEVIN DONE, AEROSPACE CORRESPONDENT

Seven UK airlines including British Airways, Virgin Atlantic, BMI
British Midland and EasyJet, on Friday took over control of the air
traffic control system, completing one of the government's most
controversial public-private partnership deals.

Completion of the National Air Traffic Services deal comes at a
critical time for the government as it tries to push through the PPP
for the London Underground.

The sale to a strategic investor of a 46 per cent stake in Nats is
the first time in Europe that management control of en route air
traffic services has passed into private hands.
```

Figure 1: Document annotated with instances from the PROTON ontology

reward was assigned to each prediction of instance — the reward was 1 if the prediction was correct, and 0 otherwise. We call this flat_F_1 since it does not consider the structure of concepts in the ontology. The other two measures were based on the BDM and LA values, respectively. As discussed in Section 3, these take into account the structure of the ontology.

	flat_F_1	BDM_F_1	LA_F_1
SVMUM	73.5	74.5	74.5
Hieron	74.7	79.2	80.0

Table 1: Comparison of the two learning algorithms Hieron and SVM with uneven margins for OBIE using three overall performance measures

Table 1 presents the experimental results for comparing the two learning algorithms SVM and Hieron. We used three measures: conventional micro-averaged flat_F_1 (%), and the two ontology-sensitive augmented F_1 (%) based respectively on the BDM and LA, BDM_F_1 and LA_F_1, which were discussed in Section 3.4. In this experiment, the International-Politics part of the OntoNews corpus was used as the test set and other two parts as the training set.

Both the BDM_F_1 and LA_F_1 are higher than the flat_F_1 for the two algorithms, reflecting the fact that the latter only counts the correct classifications, while the former two not only count the correct classifications but also the incorrect ones. However, the difference for the Hieron is more significant than that for the SVM, demonstrating an important difference between the two methods — the SVM based method just tried to learn a classifier for one concept as well as possible, while the Hieron based method not only learned a good classifier for each individual concept but also took into account the relations between the concepts in the ontology during the learning.

In terms of the conventional flat_F_1, the Hieron was slightly better than the SVM. However, if the results are measured by using the ontology-sensitive measure BDM_F_1 or LA_F_1, we can see that the Hieron performed significantly better than the SVM. Clearly, the ontology-sensitive measures such as the BDM_F_1 and LA_F_1 are more suitable than the conventional flat_F_1 to measure the performance of an ontology-dependent learning algorithm such as Hieron.

In order to analyse the difference between the three measures, Table 2 presents some examples of entities predicted incorrectly by the Hieron based learning system, their key labels, and the similarity between the key label and predicted label measured respectively by the flat measure, the BDM and the LA.

All the concepts and their relations involved in Table 2 are illustrated in Figure 2, which presents a part of the Proton ontology. This ontology section starts with the root node *Thing*, and has 10 levels of concepts with *TVCompany* as the lowest level concept. Note that the graph does not show all the child concepts for most of the nodes presented.

The conventional flat measure assigned each case a zero similarity because the examples were misclassified and the measure does not consider the structure of labels. On the other hand, both the LA and BDM take into account the structure of labels and measure the degree of a misclassification based on its position in the ontology. Hence they assign a non-zero value to a misclassification in most cases. Note that zero would be assigned in the case where the MSCA is the root node. In our experiments, all the concepts used were below the node "Entity" and so we used its immediate upper node "Thing" as root[4]. This meant that CP was always at least 1, and hence there is no zero value for BDM or LA in our experiments. This is because we consider

[4] "Thing" subsumes both "Entity" and "Property"

No.	Entity	Predicted label	Key label	Flat	BDM	LA
1	Sochi	Location	City	0.000	0.724	1.000
2	Federal Bureau of Investigation	Organization	GovernmentOrganization	0.000	0.959	1.000
3	al-Jazeera	Organization	TVCompany	0.000	0.783	1.000
4	Islamic Jihad	Company	ReligiousOrganization	0.000	0.816	0.556
5	Brazil	Object	Country	0.000	0.587	1.000
6	Senate	Company	PoliticalEntity	0.000	0.826	0.556

Table 2: Examples of entities misclassified by the Hieron based system

that if an entity's instance is recognized but with the wrong type, the system should have a non-zero reward because it at least recognized the instance in the first place. However, this could be changed according to the user's preference.

However, BDM and LA adopt different mechanisms in consideration of the ontology structure. In particular, the LA assigns the maximal value 1 if the predicted label is an ancestor concept of the key label, regardless of how far apart the two labels are within the ontological chain. In contrast, the BDM takes into account the similarity of two concepts in the ontology and assigns a distance-dependent value. The difference is demonstrated by the examples in the table. For example, in the Proton ontology, the predicted label *Organization* is the parent concept of the key label *GovernmentOrganization* in the second example, and in the third example the same predicted label *Organization* is 4 concepts away from the key label *TVCompany*, see Figure 2. Hence, the BDM value of the second example is higher than the BDM value of the third example. In the first example the predicted label *Location* is 3 concepts away from the key label *City* but its BDM value is lower than the corresponding value in the third example, mainly because the concept *Location* occupies a higher position in the Proton ontology than the concept *Organization*. Similarity is thus lower because higher concepts are semantically more general, and therefore less informative.

Another difference between the BDM and LA is that the BDM considers the concept densities around the key concept and the response concept, but the LA does not. The difference can be shown by comparing the fourth and the sixth examples. They have the same predicted label *Company*, and their key labels *ReligiousOrganization* and *PoliticalEntity* are two sub-concepts of *Organization*. Therefore, the positions of the predicted and key labels in the two examples are very similar and hence their LA values are the same. However, their BDM values are different — the BDM value of the fourth example is a bit lower than the BDM value of the sixth example. This is because the concept *PoliticalEntity* in the sixth example has two child nodes but the concept *ReligiousOrganization* in the fourth example has no child node, resulting in different averaged lengths of chains coming through the two concepts.

The BDM value in the fifth example is the lowest among the examples, mainly because the concept *Object* is in the highest position in the ontology among the examples. These differences in BDM scores show the effects of the adoption of chain density and branching factor as penalty weights in the computation of the score. These reflect the level of difficulty associated with the selection of a particular ontlogical class relative to the size of the set of candidates.

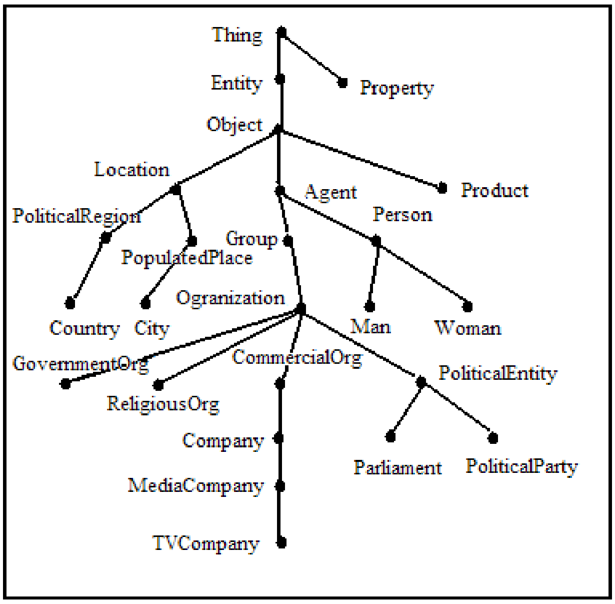


Figure 2: Subset of the Proton ontology

5. CONCLUSIONS AND FUTURE WORK

The initial observation in this paper is that binary decisions are not good enough for ontology evaluation, when hierarchies are involved. We propose an Augmented Precision and Recall measure that takes into account the ontological distance of the response to the position of the key concepts in the hierarchy. For this purpose we have developed an extended variant of Hahn's Learning Accuracy measure, called Balanced Distance Metric, and integrated this with a standard Precision and Recall metric. We have performed evaluations of these three metrics based on a gold standard corpus of news texts annotated according to the PROTON ontology, and conclude that both the BDM and LA metrics are more useful when evaluating information extraction based on a hierarchical rather than a flat structure. Furthermore, the BDM appears to perform better than the LA in that it reflects a better error analysis in certain situations.

If we compare the BDM metric with the criteria in Section 2, we observe that it conforms to the majority of these prerequisites. Although it gives an intuitively plausible score for semantic similarity on many occasions, it can be argued that in some cases it does not correlate well with human judgement. Examples 4 and 6 in Table 2 show counter-intuitively high similarity values for combinations of key and wrongly predicted labels, particularly in comparison with example 7. From a human perspective, they seem much more wrong than the erroneous classification in Example 7, and slightly more wrong than those in examples 1 and 3. This indicates a need for further tuning the BDM score with additional cost-based metrics, in order to meet human judgement criteria. In such cases, this could entail the integration of a rule which boosts similarity scores for concepts within the same ontological chain (in a more subtle way than LA), and which lowers the score for concept pairs that occur in different chains. In the future we will work on the (semi-)automatic identification of such rules.

6. ACKNOWLEDGMENTS

This work is partially supported by the EU-funded Knowledge Web network of excellence ((IST-2004-507482) and SEKT project (IST-2004-506826).

7. REFERENCES

[1] ACE. *Annotation Guidelines for Entity Detection and Tracking (EDT)*, Feb 2004. Available at http://www.ldc.upenn.edu/Projects/ACE/.

[2] E. Agirre and G. Rigau. Word sense disambiguation using conceptual density. In *Proc. of 16th International Conference on Computational Linguistics*, volume 1, pages 16–23, Copenhagen, Denmark, 1996.

[3] N. Chinchor. Muc-4 evaluation metrics. In *Proceedings of the Fourth Message Understanding Conference*, pages 22–29, 1992.

[4] P. Cimiano, S.Staab, and J. Tane. Automatic Acquisition of Taxonomies from Text: FCA meets NLP. In *Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining*, pages 10–17, Cavtat-Dubrovnik, Croatia, 2003.

[5] N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, 2000.

[6] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[7] W. Daelemans and M. O. (eds.). *Proceedings of CoNLL-2002.* 2002.

[8] O. Dekel, J. Keshet, and Y. Singer. Large Margin Hierarchical Classification. In *Proceedings of the 21st International Conference on Machine Learning (ICML-2004)*, Canada, 2004.

[9] J. G. Fiscus, G. Doddington, J. S. Garofolo, and A. Martin. Nist's 1998 topic detection and tracking evaluation (tdt2). In *Proc. of the DARPA Broadcast News Workshop*, Virginia, US, 1998.

[10] W. Frakes and R. Baeza-Yates, editors. *Information retrieval, data structures and algorithms.* Prentice Hall, New York, Englewood Cliffs, N.J., 1992.

[11] R. Garcia-Castro, D. Maynard, H. Wache, D. Foxvog, and R. G. Cabero. Specification of a methodology, general criteria and benchmark suites for benchmarking ontology tools. Technical Report D2.1.4, KnowledgeWeb Deliverable, 2005.

[12] R. Grishman and B. Sundheim. Message understanding conference - 6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, June 1996.

[13] U. Hahn and K. Schnattinger. Towards text knowledge engineering. In *Proc. of 15th National Conference on Artificial Intelligence (AAAI-98)*, pages 524–531, Menlo Park, CA, 1998. MIT Press.

[14] S. Handschuh, S. Staab, and F. Ciravegna. S-CREAM — Semi-automatic CREAtion of Metadata. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, pages 358–372, Siguenza, Spain, 2002.

[15] M. King. Living up to standards. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing*, Budapest, Hungary, 2003.

[16] A. Kiryakov, B. Popov, D. Ognyanoff, D. Manov, A. Kirilov, and M. Goranov. Semantic annotation, indexing and retrieval. *Journal of Web Semantics, ISWC 2003 Special Issue*, 1(2):671–680, 2004.

[17] Y. Li, K. Bontcheva, and H. Cunningham. SVM Based Learning System For Information Extraction. In *Proceedings of Sheffield Machine Learning Workshop*, Lecture Notes in Computer Science. Springer Verlag, 2005.

[18] Y. Li, K. Bontcheva, and H. Cunningham. Using Uneven Margins SVM and Perceptron for Information Extraction. In *Proceedings of Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 2005.

[19] Y. Li, K. Bontcheva, and H. Cunningham. Perceptron-like learning for ontology based information extraction. Technical report, University of Sheffield, Sheffield, UK, 2006.

[20] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing.* MIT press, Cambridge, MA, 1999. Supporting materials available at `http://www.sultry.arts.usyd.edu.au/fsnlp/` .

[21] D. Maynard and S. Ananiadou. Identifying terms by their family and friends. In *Proc. of 18th International Conference on Computational Linguistics (COLING)*, Saarbrücken, Germany, 2000.

[22] F. Olsson, G. Eriksson, K. Franzn, L. Asker, and P. Lidn. Notions of Correctness when Evaluating Protein Name Taggers. In *Proceedings of COLING 2002*, Taipei, Taiwan, 2002.

[23] W. Peters, N. Aswani, K. Bontcheva, and H. Cunningham. Quantitative Evaluation Tools and Corpora v1. Technical report, SEKT project deliverable D2.5.1, 2005.

[24] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. KIM – Semantic Annotation Platform. *Natural Language Engineering*, 2004.

[25] D. Roth and A. van den Bosch (eds.). *Proceedings of CoNLL-2002.* 2002.

[26] C. van Rijsbergen. *Information Retrieval.* Butterworths, London, 1979.